Variational assimilation of Lagrangian data in oceanography

Maëlle Nodet
maelle.nodet@inria.fr

April 7, 2008

Abstract

We consider the assimilation of Lagrangian data into a primitive equations circulation model of the ocean at basin scale. The Lagrangian data are positions of floats drifting at fixed depth. We aim at reconstructing the four-dimensional space-time circulation of the ocean. This problem is solved using the four-dimensional variational technique and the adjoint method. In this problem the control vector is chosen as being the initial state of the dynamical system. The observed variables, namely the positions of the floats, are expressed as a function of the control vector via a nonlinear observation operator. This method has been implemented and has the ability to reconstruct the main patterns of the oceanic circulation. Moreover it is very robust with respect to increase of time-sampling period of observations. We have run many twin experiments in order to analyze the sensitivity of our method to the number of floats, the time-sampling period and the vertical drift level. We compare also the performances of the Lagrangian method to that of the classical Eulerian one. Finally we study the impact of errors on observations.

1 Introduction

The world's oceans play a crucial role in governing the earth's weather and climate. Lack of data has been a serious problem in oceanography for a long time. Since ten years the number of observations has greatly increased,

with the availability of satellite altimeter data (ie measurements of the free-surface height of the ocean) from Geosat, Topex/Poseidon, Jason and other satellites. In addition to these remote-sensing data we have in situ data, from scientific ships, surface mooring buoys or Lagrangian drifting buoys. Among these observations, Lagrangian data, ie positions of drifting floats, play a relevant role for many reasons: firstly their horizontal coverage is very wide (the whole Atlantic Ocean, for example), secondly they give information about currents, temperature and salinity in depth which are complementary to surface information given by satellite altimeters. For these reasons many national and international programs are organized to deploy drifting floats in the world's oceans. The largest program of this type is Argo (2 055 floats on the 13th October 2005, 3 000 planned), whose floats provide also temperature and salinity profiles.

There are different types of drifting buoys. In the framework of ocean basin-scale localized experiments, oceanographers have datasets from acoustic floats. These floats emit acoustic signals which are recorded by moored listening stations, and the floats positions are calculated every six hours by triangulation. Large datasets are available especially in the Atlantic Ocean (SAMBA, ARCANE-Eurofloat, ACCE experiments). On a larger scale Argo floats are deployed in order to provide vertical temperature and salinity profiles. Assimilation of Argo thermohaline data has been successfully investigated by Forget (PhD Thesis 2004). Argo floats provide also Lagrangian information, which are their positions every ten days. Indeed they drift freely at a predetermined parking depth (around 1 000 meters), every ten days they descent to begin profiles from greater depth (2 000 meters) then they go back to the surface and they record temperature and salinity profiles during ascent. On the surface they transmit data to satellite and they are located by GPS. Thus many different floats networks and Lagrangian datasets are available.

In parallel, modeling of the ocean system has greatly improved in both quality and realism, and there are many Ocean Global Circulation Models (OGCM), like for example the OPA PArallelized Ocean model (see Madec *et al* 1998). A crucial issue for oceanographers is then to take the best advantage of different types of information included in models to one hand and in various observations to the other hand. Data Assimilation (DA) covers all theoretical and numerical mathematical methods which allow to blend as optimally as possible all sources of information (see the review by Ghil *et al* 1997 and by De Mey 1997). There are two main categories of DA methods: variational

methods based on optimal control theory (Lions 1968) and statistical ones based on optimal statistical estimation (Jazwinski 1970). Adjoint method is the prototype of variational methods, introduced in meteorology by Penenko and Obraztsov (1976). Its effective implementation in the framework of atmospheric Data Assimilation, namely four dimensional variational assimilation (4D-Var), has been studied by Le Dimet and Talagrand (1986, see also Talagrand and Courtier 1987). Introduction of 4D-Var in oceanography is even more recent (see Thacker and Long 1988, Sheinbaum and Anderson 1990). The prototype of sequential methods is the Kalman filter, introduced in oceanography by Ghil (1989) (see also the reviews by Ghil and Malanotte-Rizzoli 1991).

Assimilation of Lagrangian data is in the pipeline. Kamachi and O'Brien (1995) used the adjoint method in a Shallow-Water model with upper-layer thickness as control vector. More recently Mead (2004) has implemented a variational method based on the use of Lagrangian coordinates for Shallow-Water equations. Molcard *et al* (2003) and Özgökmen *et al* (2003) implemented optimal interpolation (which is a simple sequential method) in a reduced-gravity quasi-geostrophic model and in a primitive equations model; their method is based on conversion of Lagrangian data into velocity information. Ide, Kuznetsov, Jones and Salman (2002, 2003, 2005) used Extended and Ensemble Kalman methods to assimilate Lagrangian data into a Shallow-Water model; their method is based on an augmented state vector approach which does not require the conversion of the positions into velocity data. These teams have used simulated data in the twin experiments approach: they don't use real Lagrangian data but idealized observations simulated from a known "true state" of the ocean. Beside these studies Assenbaum and Reverdin (2005) assimilate real data available during the POMME experiment, including Argo floats data, into a very high resolution model thanks to optimal interpolation.

Previous works on Lagrangian DA were either based on sequential methods or on variational ones into very simple models. In this paper we investigate variational assimilation of drifters positions into the high resolution primitive equations model OPA.

The aim of variational assimilation methods is to identify the initial state of an evolution problem which minimizes a cost function. This cost function represents the difference between observations and their corresponding model variables. It is minimized using a gradient descent algorithm. The gradient is computed by integration of the adjoint model. Thanks to this

formulation there is no need to convert Lagrangian data into velocities data: we can use directly the position observations, although they are not variables of the ocean model, but nonlinear functions of the state variables. Moreover this method is a four dimensional one because the temporal dimension of the observations, ie their Lagrangian nature, is taken into account. The cost function involves a so-called observation operator, which links the state variables (here the velocities) and the observed data (here the positions of drifting particles). This operator is nonlinear and consequently the cost function is not necessarily convex so we used an incremental method (see Courtier *et al* 1994) in order to achieve and accelerate the minimization.

We implement our method using the primitive equations model OPA. Our configuration is an idealized wind-driven mid-latitude box model, which is representative of the different processes that are going on in the real mid-latitude ocean, as shown by Holland (1978). Then we use the twin experiments approach. As we have said before, Lagrangian observations are simulated from a known "true state", so that data are perfectly consistent with the model and it ensures there is no systematic bias in the observations. It is of course unrealistic, but twin experiments are a necessary first-step to validate our method. Indeed in this framework we know exactly the system true state and so we are able to quantify the efficiency of our method by comparing assimilated and true states. Moreover it was relevant not to use real data for many reasons: firstly, Argo floats have not been launched to provide Lagrangian information, the feasibility of exploiting their positions is absolutely not ensured. Secondly, Lagrangian datasets (from Argo to acoustic floats) are very diversified in terms of number of floats, time-sampling period of observations and drifting depth and we want to investigate the sensitivity of our method to these parameters. In order to take into account difficulties of real data (such as drift during ascent and descent for Argo floats or acoustic positioning problems) we also study the impact of errors in observations on assimilation efficiency.

The paper is organized as follows: in section 2 we describe the physical model and the Lagrangian simulated data. In section 3 we present the assimilation method and its implementation. Some numerical results are given and commented in section 4. We conclude in section 5.

2 Physical model and Lagrangian data

2.1 The Primitive Equations of the ocean

The ocean circulation model used in our study is a Primitive Equations (PE) model. These equations are derived from Navier Stokes equations (mass conservation and momentum conservation, included the Coriolis force) coupled with a state equation for water density and heat equation, under Boussinesq and hydrostatic approximations (for more details see Lions *et al* 1992 and Temam and Ziane 2004).
These equations are written as

$$\begin{cases} \partial_t u - b\Delta u + (U.\nabla_2)u + w\partial_z u - av + \partial_x p = 0 & \text{in } \Omega \times (0,t_f) \\ \partial_t v - b\Delta v + (U.\nabla_2)v + w\partial_z v + au + \partial_y p = 0 & \\ \partial_z p - gT = 0 & \\ \partial_t T - b\Delta T + (U.\nabla_2)T + w\partial_z T + fw = 0 & \text{in } \Omega \times (0,t_f) \\ w(x,y,z) = -\int_0^z \partial_x u(x,y,z') + \partial_y v(x,y,z')\, dz' & \text{in } \Omega \times (0,t_f) \\ U(t=0) = U_0, \qquad T(t=0) = T_0 & \text{in } \Omega \end{cases} \tag{1}$$

where

- $\Omega = \Omega_2 \times (0,1)$ is the circulation basin, where Ω_2 is a regular bounded open subset of $\mathbb{R}^2$, x and y are the horizontal variables and $z \in (0,1)$ is the vertical one, $(0,t_f)$ is the time interval;

- $U = (u,v)$ is the horizontal velocity, w is the vertical velocity, T the temperature and p the pressure;

- $U_0 = (u_0, v_0)$ and T_0 the initial conditions;

- $(\nabla_2.)$ is the horizontal divergence operator and $\Delta = \partial_{xx} + \partial_{yy} + \partial_{zz}$ the 3-D Laplace operator;

- a, b, f, g are physical constants.

The space boundary conditions are

$$\begin{cases} \partial_z u = \tau_u, \quad \partial_z v = \tau_v, \quad T = 0 & \text{on } \Gamma_t \\ u = 0, \qquad v = 0, \qquad T = 0 & \text{on } \partial\Omega \setminus \Gamma_t \\ \int_{z=0}^{1} \partial_x u + \partial_y v\, dz = 0 & \text{in } \Omega_2 \end{cases} \tag{2}$$

where $\tau = (\tau_u, \tau_v)$ is the stationary wind-forcing, $\partial\Omega$ is the boundary of Ω and $\Gamma_t = \Omega_2 \times \{z = 1\}$ is its top boundary.

2.2 Model and configuration

We are using the OPA ocean circulation model developed by LODYC (see Madec *et al* 1998), in its 8.1 version. OPA is a flexible model and can be used either in regional or in global ocean configuration. The prognostic variables are the three-dimensional velocity field (u, v, w) and the thermohaline variables T and S. Discretization is based on finite differences in space and time (leap-frog scheme in time). Various physical choices are available to describe ocean physics.

The characteristics of our configuration are as follows:

– The domain is $\Omega = (0, l) \times (0, L) \times (0, H)$ (longitude, latitude, depth), with $l = 2800$ km, $L = 3600$ km and $H = 5000$ m. It extends from -56^o to -24^o West longitude, and from 22.5^o to 47.5^o North latitude.

– The horizontal resolution is 20 km, there are 11 vertical levels, so that the number of grid points is $180 \times 140 \times 11 = 277200$.

– The time step is 1200 seconds.

– The model is purely wind-driven.

This configuration is a classical eddy-resolving double-gyre circulation. As shown by Holland (1978) this model is representative of real mid-latitude oceans, where circulation is highly nonlinear, non-stationary and where oceanic turbulence is very active. Indeed a very active and unstable mid-latitude jet develops at the convergence of the subpolar gyre and the subtropical gyre. Non-stationary mesoscale eddies form also along the jet. So this model shows dynamically different processes such as large-scale gyres, mid-latitude jet, mesoscale eddies and also western boundary currents which interact in a complex way. Therefore this configuration is a difficult and interesting situation to study Lagrangian DA.

The model is integrated for 25 years until a statistically steady-state is reached, which is our "true state" for the twin experiments. Figure 1 shows an instantaneous horizontal velocity field at the surface, on the whole horizontal grid on the left and on a reduced grid on the right. We can see the mid-latitude jet and some mesoscale eddies.

2.3 Lagrangian data

Lagrangian data are positions of drifting floats. These floats drift between $z_0 - a$ and $z_0 + a$ where z_0 is given by the user and a is around 25 meters, so that we can assume that the floats drift at fixed depth z_0, ie in the horizontal

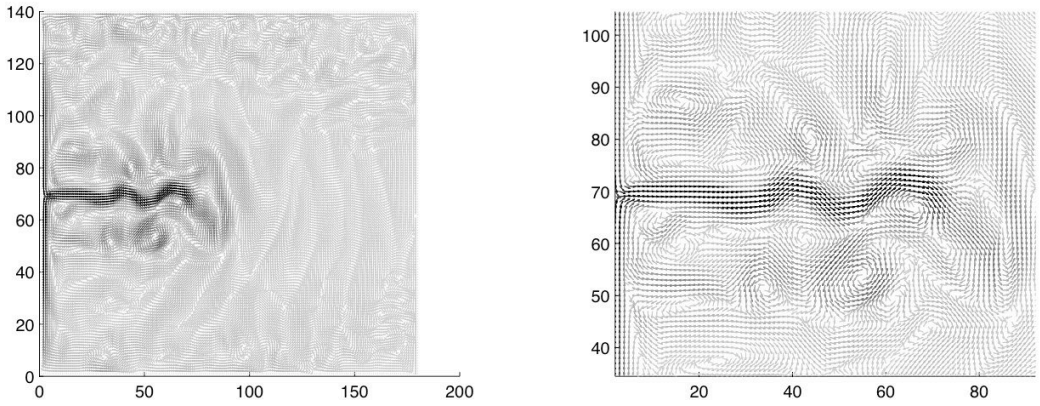

Figure 1: Instantaneous horizontal velocity field of the true state. On the left the velocity field at the surface on the whole horizontal grid. On the right the velocity field at the surface on a reduced grid centered on the mid-latitude jet. Dark grey vectors represent large velocities.

plane $z = z_0$ (Assenbaum 2005, personal communication). We denote by $\xi(t) = (\xi_1, \xi_2)(t)$ the position of one float at time t in the plane $z = z_0$. $\xi(t)$ is the solution of the following differential equation:

$$\left\{ \begin{array}{rcl} \dfrac{d\xi}{dt} & = & U(t, \xi(t), z_0) \\ \xi(0) & = & \xi_0 \end{array} \right. \tag{3}$$

where $U = (u, v)$ is the horizontal velocity of the flow and ξ_0 the initial position of the float. It is important to notice that the mapping $U \mapsto \xi$, which links the variables of the model and the Lagrangian observations is nonlinear. In the twin experiments approach, observations are simulated by the model. The true initial state of the ocean is given and OPA model computes the true velocities of the ocean during a ten-day window. We compute on-line perfect observations.
To do that we integrate numerically the equation (3) using a leapfrog scheme. This requires the velocity U along the trajectory of the float (ie out of the grid). To achieve this we use the following continuous 2D interpolation

'interp$(U,(x,y))$' of the vector field U at the point (x,y):

$$\begin{array}{ll} x_1 = \lfloor x \rfloor, & y_1 = \lfloor y \rfloor, \\ u_1 = U(x_1, y_1), & u_2 = U(x_1+1, y_1), \\ u_3 = U(x_1, y_1+1), & u_4 = U(x_1+1, y_1+1), \\ \multicolumn{2}{l}{\text{interp}(U,(x,y)) = u_1 + (u_2-u_1)(x-x_1) + (u_3-u_1)(y-y_1)} \\ \multicolumn{2}{l}{\qquad +(u_1-u_2-u_3+u_4)(x-x_1)(y-y_1)} \end{array}$$

where $\lfloor . \rfloor$ denotes the floor function, (x_1, y_1), (x_1+1, y_1), (x_1, y_1+1) and (x_1+1, y_1+1) are the grid points which are the nearest neighbors to (x,y). This function is piecewise affine with respect to x and y, continuous with respect to (x,y), linear with respect to u. Thus it is not differentiable in (x,y) everywhere. More precisely it is not differentiable at (x,y) if and only if $x = x_1$ or $y = y_1$. It will be a problem to derive the adjoint code, see paragraph 3.3. However it is accurate enough to approximate the solution of equation (3) and it is very costly to use a differentiable interpolation. Indeed such a method (like cubic splines for example) would compute each interpolated value from the whole field u (ie the values of u at every horizontal grid point) and we would have to inverse a n by n matrix (where $n = 25\,200$ is the number of horizontal grid points) at every time step and this is not workable.

Let us denote $\xi_k = (\xi_{1,k}, \xi_{2,k})$ the horizontal position of the float at time t_k, U the horizontal velocity of the fluid at time t_k, U_k the velocity at point ξ_k, and h the time step of the ocean model. The algorithm step is schematically

$$\left\{ \begin{array}{rcl} \xi_k & = & \xi_{k-2} + 2h\, U_{k-1} \\ U_k & = & \text{interp}(U, \xi_k) \end{array} \right.$$

The dataset is $\{\xi_N, \xi_{2N}, \xi_{3N} ...\}$, where N is an integer. The duration between two data is thus the product of N by the time-step h of the code ($h = 1200$ seconds). We call this the time-sampling period. For example if $N = 72$ we have one data per float and per day and the time-sampling period is thus one day.

In order to simulate real floats we can add errors to the simulated observations. Origins of errors are multiple: for acoustic floats, inaccuracy can come from acoustic sources (accuracy of their positioning, clocks accuracy, bottom topography – acoustic shadow problem, etc.), floats (listening period accuracy, complexity of the trajectory, technical problem – temporary "deafness", etc.) or communications quality. For Argo floats errors are due

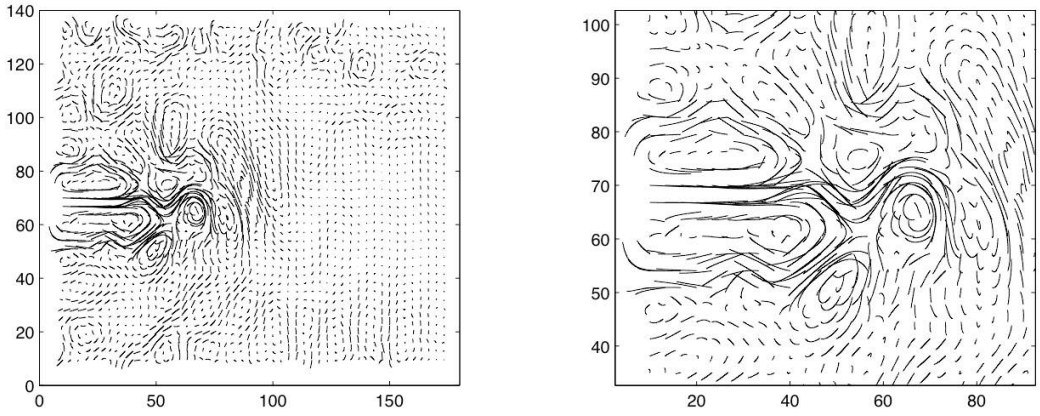

Figure 2: Trajectories of 2 000 floats drifting at level 4 during ten days. On the left, trajectories on the whole horizontal grid at level 4. On the right, trajectories on a reduced grid around the mid-latitude jet. Very different trajectories are observed: short to long, half-circle or straight.

to drift during ascent and descent and also to drift at the surface between ascent/descent and satellite communication. Errors amplitude is around 3 to 4 kilometers for acoustic floats (T. Reynaud, private communication) and 2 to 6 kilometers for Argo floats (M. Assenbaum, private communication). Figure 2 represents perfect data simulated by the algorithm with 2 000 floats for 10 days at level 4 (1 000 meters).

3 Description and implementation of the variational assimilation method

3.1 Description of the assimilation problem

Without loss of generality we assume that there is only one assimilated data: the dataset is the position $\xi(t_1) = (\xi_1(t_1), \xi_2(t_1))$ in the horizontal plane $z = z_0$ of a single float at a single time t_1. We use here the notations established by Ide *et al* (1997). We denote by $\mathbf{y}^o = (\xi_1(t_1), \xi_2(t_1))$ this data. Our problem is to minimize the following cost function with respect to the

control vector $\mathbf{x}$:

$$\begin{array}{rcccc} \mathcal{J}(\mathbf{x}) & = & \frac{1}{2}\|\mathcal{G}\mathcal{M}(\mathbf{x}) - \mathbf{y}^o\|^2 & + & \frac{\omega}{2}\|\mathbf{x} - \mathbf{x}^b\|^2_{\mathbf{B}} \\ & = & \mathcal{J}^o(\mathbf{x}) & + & \omega\,\mathcal{J}^b(\mathbf{x}) \end{array} \tag{4}$$

where

– the control vector $\mathbf{x} = (u_0, v_0, T_0)$ is the initial state vector,

– the $\mathbf{B}$-norm is calculated thanks to the background error covariance matrix $\mathbf{B}^{-1}$: $\|\psi\|^2_{\mathbf{B}} = \psi^T\mathbf{B}^{-1}\psi$,

– $\mathbf{x}^b$ is another initial state of the ocean, called the background or first guess, which is required to be close to the minimum $\bar{\mathbf{x}}$,

– $\mathcal{M}$ is the discrete ocean model and $\mathcal{M}(\mathbf{x})$ is the discrete state vector (one value per variable, per grid-point and per time step),

– $\mathcal{G}$ is the discrete nonlinear observation operator, which links the state of the fluid (and especially the horizontal velocity U) with the data, $\mathcal{G}\mathcal{M}(\mathbf{x}) = \xi(t)$ where $\xi(t)$ is defined by equation (3),

– $\|.\|$ is the euclidean norm in $\mathbb{R}^2$,

– $\mathbf{y}^o$ is the observations vector.

Then $\mathcal{J}^o$ quantifies the misfit between observations and the state of the system, $\mathcal{J}^b$ represents the distance (in terms of the $\mathbf{B}$-norm) between the control vector and the background. It is also a regularization term thanks to which the inverse problem of finding the minimum $\mathbf{x}^*$ becomes well-posed. The parameter ω represents the relative weight of the regularization term with respect to the observation term and it must be chosen carefully.

3.2 Numerical variational assimilation : incremental 4D-Var

Four dimensional variational assimilation (4D-Var, see Le Dimet and Talagrand 1986) is an iterative numerical method which aims to approximate the solution $\mathbf{x}^*$ of discrete assimilation problems with cost function of type (4). In 4D-Var a gradient descent algorithm is used to minimize the cost function, the gradient being obtained by solving the discrete adjoint equations. It is an efficient method but it is very costly when the direct model and the observation operator are not linear, for at least two reasons. Firstly every iteration of the adjoint method requires one integration of the full non linear direct model and one integration of the adjoint of the linearized model. Secondly the cost function is not necessarily convex and the minimization process may

converge to a local minimum, or it may take considerable time to converge, or may not converge at all.
Incremental 4D-Var (see Courtier *et al* 1994) avoids, to some extent, both of these problems. In this approach, the nonlinear model is approximated by a simplified linear model (called tangent linear model) and the nonlinear observation operator is linearized around a reference state. The cost function becomes quadratical, it has a unique minimum and this minimum is assumed to be close to the one of the full non quadratical cost function. In that case the minimization process converges quickly. Moreover this approach takes into account weak nonlinearities, because the tangent linear model and the adjoint model are updated three or four times.

Remark 1 *The approximation of the full nonlinear model by the tangent linear model is called the* tangent linear hypothesis *(TLH). In our highly nonlinear configuration, we have to use a ten-day time-window so that the TLH is valid (see section 4).*

3.3 Implementation in OPAVAR

The OPAVAR 8.1 package developed by Weaver *et al* (2003) includes the direct non linear model OPA 8.1 developed by LODYC, the tangent linear model, the adjoint model and a minimization module. Weaver has implemented a preconditioning through the $\mathbf{B}$ matrix, via the change of variables $\delta\mathbf{w} = \mathbf{B}^{-1/2}\delta\mathbf{x}$, following the method introduced by Courtier *et al* (1994). The observation operators of OPAVAR 8.1 are interpolation and projection operators. To assimilate Lagrangian data we have implemented the non linear observation operator (see section 2.3), its linearization around the reference trajectory and the adjoint of the linear observation operator.
To obtain the tangent (and adjoint) codes of the discrete observation operator we use the recipes for (hand-coding) adjoint code construction of Talagrand (1991) and Giering and Kaminski (1998). The direct and tangent algorithms are schematically:

- Direct code:
$$\left\{ \begin{array}{l} \xi_k = \xi_{k-2} + 2h\, U_{k-1} \\ U_k = \mathrm{interp}(U, \xi_k) \end{array} \right.$$
where 'interp' is the interpolation function of U at point ξ (see section 2.3).

- Linear tangent code:

$$\begin{cases} \delta\xi_k = \delta\xi_{k-2} + 2h\,\delta U_{k-1} \\ \delta U_k = \text{interp}(\delta U, \xi_k) + \delta\xi_k.\partial_{(x,y)}\text{interp}(U, \xi_k) \end{cases}$$

where '$\partial_{(x,y)}$interp' is the derivative of the 'interp' function with respect to (x, y). The term $\partial_{(x,y)}$interp is specific to Lagrangian data. It leads to a slight difficulty, because the function 'interp' is linear with respect of U but it is not derivable in (x, y) at points with integer coordinates. Thus we have chosen the values of that derivative at these points, using finite centered differences.

4 Numerical results

In this section we present the results of our numerical experiments. We begin with a brief description of our choices.

Background and time-window width. In these experiments we have assimilated only Lagrangian data and we assume that the true initial temperature and salinity were known. Background and time-window width are related because of the incremental formulation: indeed the full nonlinear model is linearized *around the background over the whole time-window*. When the background is too different from the true state or the time-window is too wide, approximation errors are large ie the *tangent linear hypothesis* (TLH) is not valid any more. So we compute some correlations to choose both of them. If we denote $\mathbf{x}^t$ the true state and $\mathbf{M}$ the tangent linear model, we can compute the nonlinear and linear perturbations δ_1 and δ_2:

$$\delta_1 = \mathcal{M}(\mathbf{x}^t) - \mathcal{M}(\mathbf{x}^b), \quad \delta_2 = \mathbf{M}(\mathbf{x}^t - \mathbf{x}^b)$$

Then we compute (as a function of time) the spatial correlation between δ_1 and δ_2 according to the formula:

$$Cor(\delta_1, \delta_2) = \frac{\langle \delta_1 \delta_2 \rangle - \langle \delta_1 \rangle \langle \delta_2 \rangle}{\sqrt{(\langle \delta_1^2 \rangle - \langle \delta_1 \rangle^2)(\langle \delta_2^2 \rangle - \langle \delta_2 \rangle^2)}} \tag{5}$$

where $\langle X \rangle$ is the spatial mean of X. The closer to 1 the correlation is, the better the adequacy between the fields is. Table 1 shows correlation at the end of the time-window between linear and nonlinear perturbations

Table 1: Background and time-window width choices: spatial correlation between nonlinear and linear perturbations at the end of the time-window, according to formula (5). The background is the state of the ocean 10 days or 1 month before the true initial state.

time-window width	background	correlation
10 days	10 days	0.80
10 days	1 month	0.67
20 days	10 days	0.50
20 days	1 month	0.42

for different time-window widths (10 or 20 days) and different background choices (the state of the ocean 10 days or 1 month before the true initial state). We can see that the TLH is not valid with a 20-day window. In the sequel we use a 10-day time-window and the state of the ocean ten days before the true one as a background state, as in this context the TLH is valid. We ran the model with the background as initial state and we obtained a "without-assimilation" state, called background in the sequel. It will be compared to the assimilated state in order to quantify the efficiency of the assimilation process.

The B matrix. The choice of the **B** matrix is crucial because of its dual purpose (preconditioning and regularization). Firstly we have tested very simple matrices (identity, energy weights) and the results were quite bad: the convergence was very slow and the analysis increments were very noisy. These matrices are used in OPAVAR only for debugging purpose, with extremely idealized assimilation context, for example when the whole state vector is observed ie with data everywhere. In order to obtain smoother increments and to accelerate the convergence, we used then the diffusion filter method (see Weaver and Courtier 2001) which gives good results.

Diagnostics. Our diagnostics are based on RMS error between the true velocity and the assimilated one, compared with the RMS error between the true velocity and the background one. The RMS error is plotted as a function of time or of the vertical level or of another parameter. For example, we have

the following formula for the time-dependent RMS error:

$$\text{error}(u,t) = \Big(\frac{\sum_{i,j,k} |u_t(i,j,k,t) - u(i,j,k,t)|^2}{\sum_{i,j,k} |u_t(i,j,k,t)|^2}\Big)^{1/2} \tag{6}$$

where u_t is the true state, u the assimilated state (or the background), t is the time and (i,j,k) a grid point, where (i,j) are the horizontal coordinates and k the vertical one.
We made the following experiments: first experiment and diagnostics in section 4.1, sensitivity to the floats network parameters (time sampling of the position measurements, number of floats, drifting level, coupled impact of number and time sampling) in section 4.2, comparison with another variational method in section 4.3.

4.1 First experiment

We present here the results of a typical experiment. There are 3 000 floats drifting at level 4 in the ocean for 10 days. The Lagrangian data are collected once a day. Thus the total amount of data is $2 \times 3\,000 \times 10 = 60\,000$.
Figure 3 (on the left) shows the RMS error of the experiment as function of time, according to formula (6). We have put the error for the background (= without assimilation state) on the same plot.
Figure 3 (on the right) shows the total RMS error as a function of the vertical level (where 1 represents the surface and 10 the bottom), according to the formula:

$$\text{error}(u,k) = \Big(\frac{\sum_{i,j,t} |u_t(i,j,k,t) - u(i,j,k,t)|^2}{\sum_{i,j,t} |u_t(i,j,k,t)|^2}\Big)^{1/2} \tag{7}$$

We can see that the error with assimilation is twice lower than without. Moreover the assimilation process improves every vertical level and not only the 4th one.

Remark 2 *We can notice that the RMS error at the beginning are quite large. It is explained by the following fact: the relative weight of the regularization term $\mathcal{J}^b$ (see section 3.1) has to be large enough to ensure the convergence of the minimization process, thus the assimilated state is a compromise between background and observations.*

Figure 4 shows the horizontal velocity field $U = (u, v)$ at level 1 at the final time. We can notice that the main patterns as the mid-latitude jet and the bigger eddies are quite similar to the true ones.

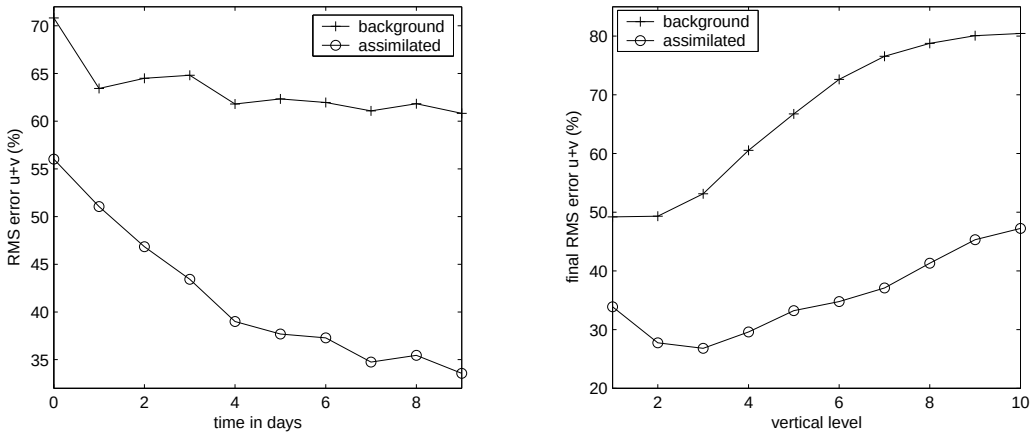

Figure 3: First experiment: u+v RMS errors corresponding to the assimilation of the positions (sampled 4 times a day) of 3 000 floats drifting at level 4. On the left, RMS error as a function of time. On the right, RMS error as a function of the vertical level. For reference the error without assimilation (background) is also displayed.

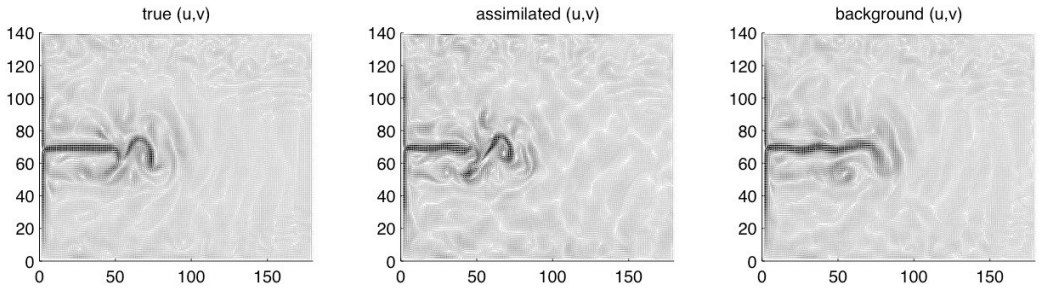

Figure 4: First experiment: horizontal surface velocity field at the final time. The true field is displayed on the left. The field corresponding to the assimilation of floats positions is displayed on the middle. For reference the field obtained without assimilation (background) is displayed on the right.

Table 2: RMS errors for u+v (in %) at different instants, during a 30-day experiment, with and without assimilation of the positions of 1 000 floats drifting at level 4, sampled once a day.

Experiment	t = 0	t = 9 days	t = 19 days	t = 29 days
Without Assimilation	70.8	61.8	62.3	59.8
With Assimilation	55.4	34.9	21.7	13.6

Longer experiments. We have seen before that the TLH is not valid for a 20-day window, so that we cannot use incremental 4D-Var for longer windows. However we can restart the assimilation process over the next 10-day window in order to do longer experiments: the new background (at the beginning of the new window) is the previous assimilated state (at the end of the previous window), so that the new background carries information from the previous assimilation process. Table 2 shows the relative RMS errors (7) for $u + v$ of a 30-day experiment at different times. We can see that error with assimilation is lower than 15% at the end of the window, ie it is less than one fourth of the error without assimilation. This is a very good result: 3 succesive assimilation processes enable to reconstruct a very good approximation of the true state.

4.2 Sensitivity to the floats network parameters

In operational oceanography, sensitivity analysis is central to the observational network design. Indeed, in situ and remote observation instruments are very expensive (e.g. one Argo floats costs 15 000$) and they must be optimally used. Ngodock (PhD thesis 1996) shows that second order analysis (ie the derivation of the optimality system or *second order adjoint system*, see also Wang *et al* 1992) enables to analyze the sensitivity of the 4D-Var assimilation system to the design of the observational network. Second order adjoint information (see also the review paper by Le Dimet *et al* 2002) is actually central to adaptive observation network and observation targeting issues. However it requires the storage of model, tangent and adjoint trajectories, so that it is not workable in OPAVAR at the present time because of computer memory limitations.

So we have performed a lot of experiments to analyze the sensitivity to various parameters of our assimilation process. Indeed in the ocean the network's

parameters can widely change, from Argo floats (1000-2000 meters depth, one data per 10 days) to acoustic floats (various depth, time sampling period around 6 hours) or drifters in the upper ocean (near surface, time sampling period very short)...
Here are the parameters that we consider:

- the time sampling period, varying from 6 hours to 10 days,
- the number of floats, varying from 300 to 3000,
- the vertical level of drift, varying from 1 (surface) to 10 (bottom),

We analyze also the coupled effect of the number and the time sampling period.

4.2.1 Sensitivity to the time sampling period.

The framework of this experiment is the following: we performed seven different experiments with exactly the same initial conditions, namely 3 000 floats at level 4. The only difference in these experiments is the time sampling period, which will be denoted shortly by TSP in the sequel. The experiments are denoted by TSP-xxx where xxx is the time sampling period, in hours (6 or 12h) or in days (1, 2, 3, 5 or 10d). Figure 5 shows the RMS error as a function of time for each TSP experiment, except (for readability) experiments TSP-6h, TSP-12h and TSP-2d. It shows also the total RMS error as a function of the time sampling period. We can see that our method is robust with respect to the increase of the time sampling period. This is very encouraging. Indeed every prior study is very sensitive to the TSP and shows quite bad results when the TSP is larger than 2 or 3 days (see Molcard *et al* 2003, Mead 2005 and section 4.3). Our method does not show this sensitivity, velocities are quite well reconstructed even when the TSP is large and especially with a 10-day period, which is a very positive result.

4.2.2 Sensitivity to the number of floats.

We perform five experiments with varying number of floats drifting at the same vertical level (4) and with positions sampled with the same period (6 hours). The experiments are denoted by NUM-xxx, where xxx is the number of floats. Figure 6 shows the RMS error as a function of time and the total RMS error as a function of the number of floats for each experiment. We can see that the number of floats has great influence on the results. Under a minimal number (1 000) the velocities are badly reconstructed, undoubtedly

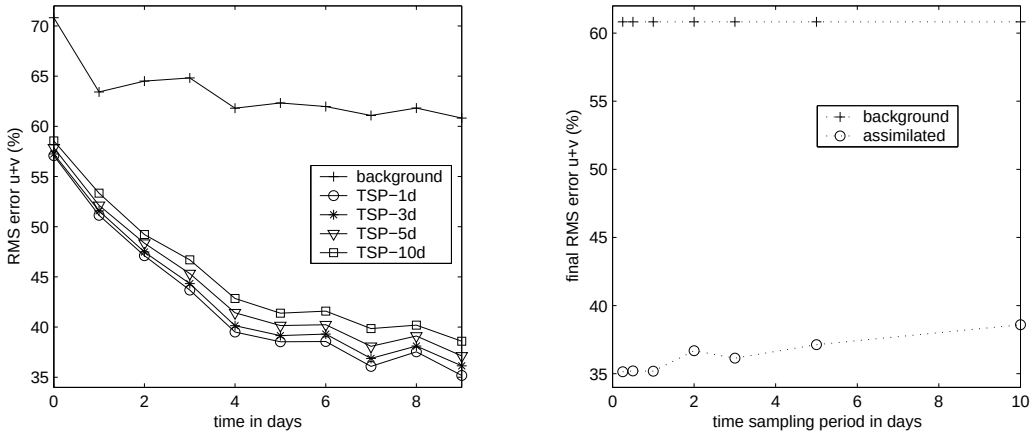

Figure 5: Sensitivity to the time sampling period: u+v RMS errors corresponding to assimilation of 3 000 floats' positions with different time-sampling periods of observation. On the left: error as a function of time for each 'TSP' experiment and for the background (without assimilation reference state). On the right: final error as a function of the TSP; the error without assimilation is also displayed.

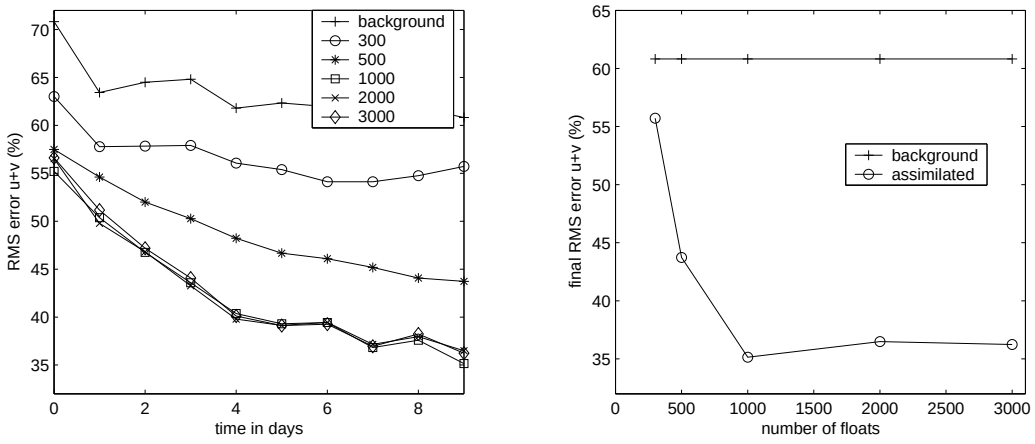

Figure 6: Sensitivity to the number of floats: u+v RMS errors corresponding to assimilation of 300 to 3 000 floats' positions. On the left: error as a function of time for each 'NUM' experiment and for the background (without assimilation). On the right: final error as a function of the number of floats; the error without assimilation is also displayed.

because there is not enough information to constrain the flow. However, when we perform longer experiments, we get satisfactory results for small numbers like 500 and 300. For a ten-day window the results are optimal with a 1 000 floats network and they don't improve with higher numbers. Obviously the information becomes redundant and it is useless to add floats. The associated density is one float per 10 000 km^2, which is ten times more than the planned Argo density (namely around 100 floats in our configuration). Even if we perform longer experiments, the Argo density is too small to constrain the velocity field. It is more appropriate to use Lagrangian data from localized experiments such as acoustic floats launchings in the Atlantic Ocean (like e.g. SAMBA), whose floats densities are higher.

4.2.3 Sensitivity to the vertical drift level.

Again we perform seven experiments with 3 000 floats drifting at varying vertical level and fixed TSP (6 hours). As usually we denote by LEV-x the experiment involving floats at level x. Figure 7 shows the RMS error as a function of time for upper levels on the left and lower levels on the right.

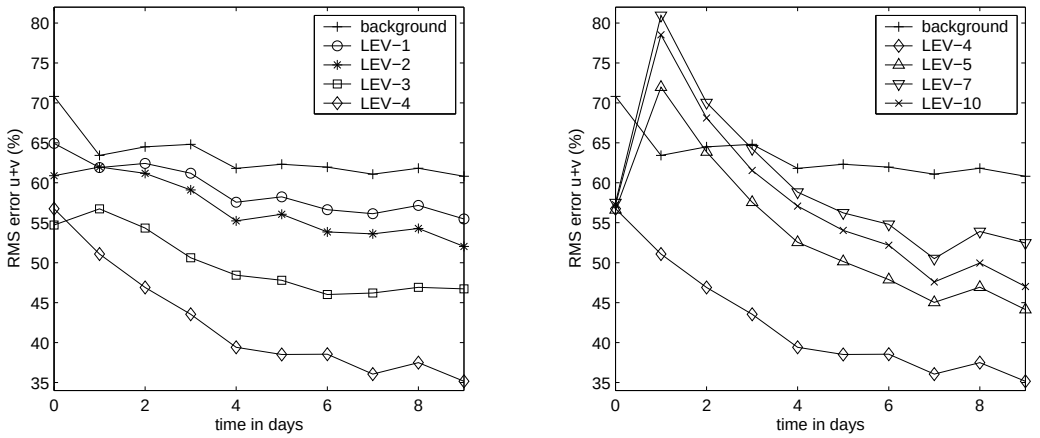

Figure 7: Sensitivity to the vertical drift level: time-evolution of the u+v RMS errors corresponding to the assimilation of positions of floats drifting at different depths. On the left: error for upper levels experiments (above 1 000 meters) and for the background (without assimilation). On the right: error for lower levels experiments (below 1 000 meters) and for the background (without assimilation).

Again the results are very sensitive to the position of the floats. The three best levels are 3, 4 and 5, ie the intermediate levels. From a physical point of view it is coherent because the information propagates vertically with a finite velocity so that very upper (1, 2) and very lower (7 to 10) levels are penalized. Moreover upper levels (1 to 4) are the most energetic ones (from the kinetic turbulent energy point of view), quasi ten times more than the lower ones (levels 5 to 10), it seems quite natural that the best results are obtained with floats drifting at level 4 which is both intermediate and energetic.

4.2.4 Coupled impact of number of floats and time sampling period.

Here we look at the coupled effect of varying number of floats and varying TSP, for example in order to answer the following question: is the total number of data an important variable to measure the efficiency of the assimilation? So we perform nine experiments, denoted by nnn-xxx where nnn is the number of floats and xxx is the TSP. These experiments and their final

Table 3: Coupled impact of number of floats and TSP: Final RMS error corresponding to assimilation experiments with 500 to 2 000 floats and positions sampled every 1 to 5 days. Total number of observations is also given. The "background experiment" results are also shown for reference.

Experiment	Total number of data	Final Error (%)
500-5D	1 000	46.6
500-3D	1 500	44.0
500-1D	5 000	44.0
2000-5D	4 000	37.1
2000-3D	6 000	36.8
2000-1D	20 000	35.9
1000-5D	2 000	35.1
1000-3D	3 000	35.1
1000-1D	10 000	34.8
Background	no data	60.8

RMS error are described in Table 3. Figure 8 represent the RMS error as a function of time for the 500-xxx experiments on the left, 1000-xxx in the middle and 2000-xxx on the right with the same scale on the axis of ordinates. The results are complementary to the precedent experiments. Indeed we can see that 1 000 is an optimal number for this configuration whatever the TSP and that our method is stable with respect to large TSP whatever the number of floats. Thus we can conclude that, in our configuration, it seems optimal to launch around 1 000 floats and that the TSP can be chosen quite large.

4.3 Comparison with the "Eulerian" method

A classical method in oceanography is to assimilate the velocity observations deduced from the Lagrangian data according to the following finite differences formula:

$$\begin{array}{rcl} \dfrac{\xi_1(t_{k+1}) - \xi_1(t_k)}{t_{k+1} - t_k} & \approx & u(\xi_1(t_k), \xi_2(t_k), z_0, t_k) \\ \dfrac{\xi_2(t_{k+1}) - \xi_2(t_k)}{t_{k+1} - t_k} & \approx & v(\xi_1(t_k), \xi_2(t_k), z_0, t_k) \end{array} \tag{8}$$

Then the velocity data are treated as Eulerian data (measured at non-fixed points). We implement this method in the 4D-Var framework. The obser-

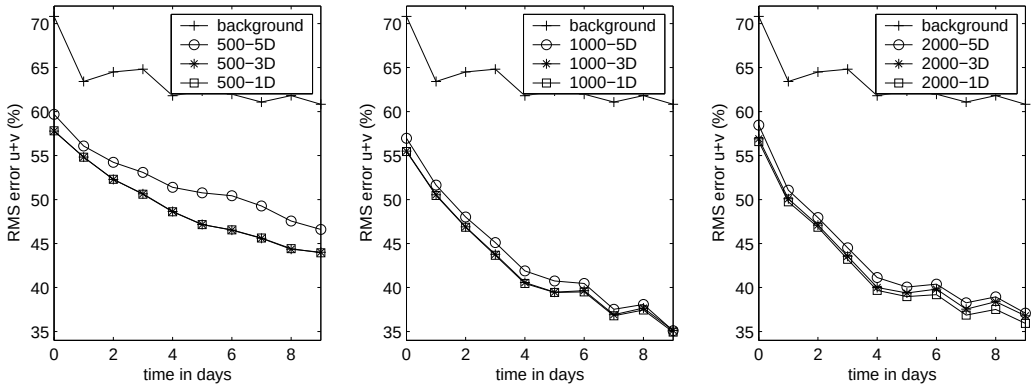

Figure 8: Coupled impact of number and time sampling period: time-evolution of the u+v RMS errors corresponding to assimilation experiments with 500 to 2 000 floats and positions sampled every 1 to 5 days: experiment with 500 floats on the left, 1 000 in the middle and 2 000 on the right. For reference, the background error is also displayed on each plot.

vation operator is much easier to write (and to differentiate and transpose) because it is an interpolation at the points of the true (fixed) floats trajectories. We compare the results for this method said "Eulerian" and for our "Lagrangian" one. Experiments have the same characteristics (3 000 floats and varying TSP), their names are LAG-xxx or EUL-xxx with xxx the TSP, where xxx is the TSP.

Figures 9 and 10 represent the RMS error as a function of time and vertical level. We can see that the "Eulerian" approach is slightly better than the "Lagrangian" one when the TSP is small (one day), moreover its computation time is 10% lower. Indeed the TSP is small enough so that the formula (8) is a very good approximation: the displacement vector between two successive positions is quasi tangent to the trajectory (ie quasi collinear with the velocity vector). However we can see that the error for the "Lagrangian" method is more homogeneous as a function of the vertical level. For larger TSP (3 days or more) the "Lagrangian" method is obviously better than the "Eulerian" one: the approximation formula (8) is not valid any more. We can see that our method is able to extract information from the positions data even if the TSP is large.

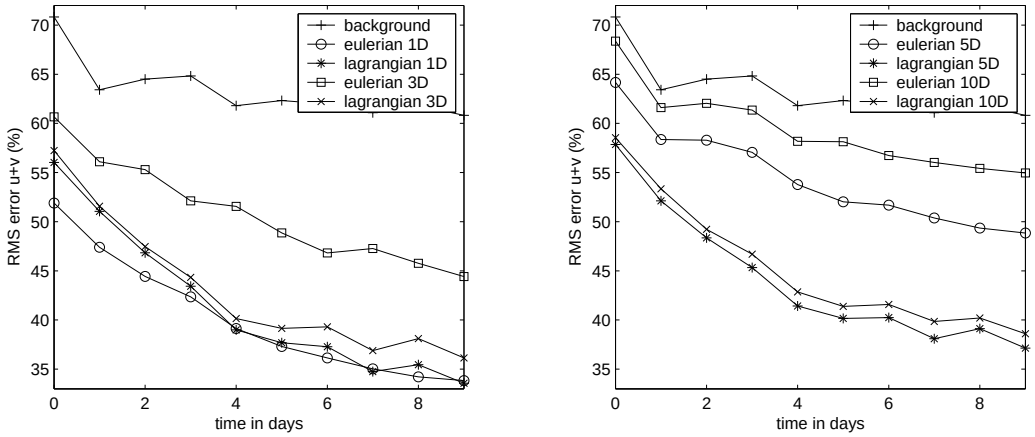

Figure 9: Comparison Eulerian/Lagrangian: time-evolution of the RMS error for u+v corresponding to the assimilation of 3 000 floats with different TSP. On the left, errors for the Lagrangian and Eulerian methods with small TSP. On the right, errors for the Lagrangian and Eulerian methods with large TSP. For reference, the background error is also displayed on each plot.

4.4 Assimilation of noisy observations

In order to deal with real data issues, a necessary first-step is to study the impact of observation errors in the twin experiments framework. To do that, we simulate as previously perfect data from the "true state" with 1 000 floats drifting at level 4, their positions being sampled once a day. Then we add a random Gaussian noise to the computed positions. In the sequel, the word "error" represents the amplitude of the noise. As we said before, real errors are about 2 to 6 kilometers. However our system is idealized so we study the impact of errors up to 20 kilometers. The total displacement of one float between initial and final positions is around 25 kilometers (in steady regions) to 90 km (in the mid-latitude jet region), so that a 10 to 20-km noise is significant for most of the floats.
Figure 11 represents RMS errors as a function of time (on the left) for experiments with 0 to 20km errors and for the background (without assimilation). On the right we plot RMS errors as a function of observation error amplitude. The RMS error is very stable with increasing noise amplitude: our method is able to extract information even when the error amplitude is not negligible with respect to floats displacement.

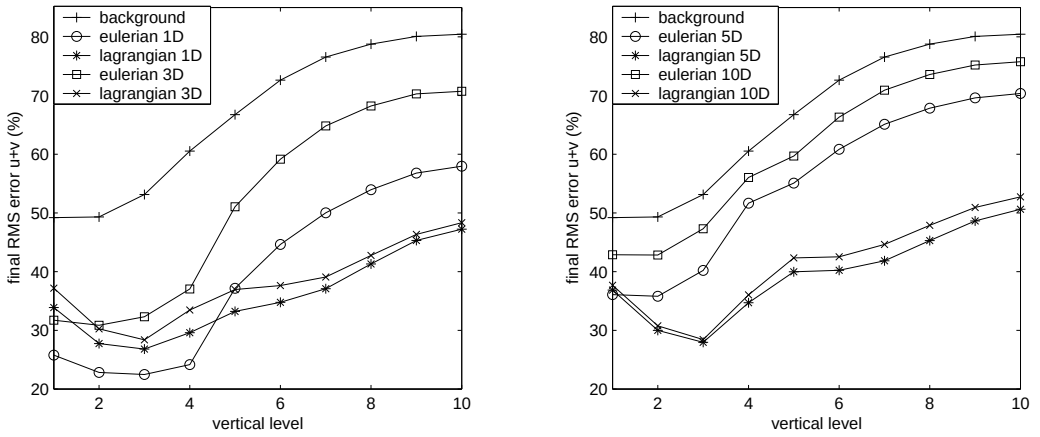

Figure 10: Comparison Eulerian/Lagrangian: final RMS error for u+v as a function of the vertical level, corresponding to the assimilation of 3 000 floats with different TSP. On the left, errors for the Lagrangian and Eulerian methods with small TSP. On the right, errors for the Lagrangian and Eulerian methods with large TSP. For reference, the background error is also displayed on each plot.

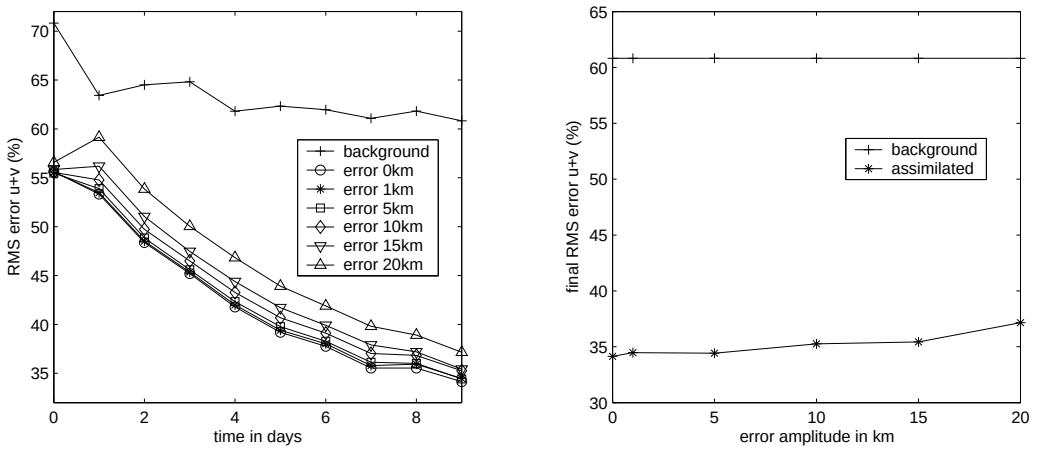

Figure 11: Impact of observation errors: u+v RMS errors corresponding to assimilation of noisy observations. On the left: error as a function of time for experiments with noise amplitude from 1km to 20km. For reference, results for experiments without noise (0km) and without assimilation (background) are also displayed. On the right: final error as a function of the amplitude of the noise; the error without assimilation is also displayed.

Figure 12 shows the evolution of the cost function value and its gradient during the assimilation process. The abscissa represents the total number of iterations. Here we perform four outer loops: in each outer loop we perform ten inner minimization loops, in which the cost function is minimized, as we can see on the left. We can see that the gradient norm decreases and the cost function converges even in the presence of noise in data.

5 Conclusion

This paper shows that the problem of assimilating Lagrangian data can be solved by a variational adjoint method into a realistic primitive equations ocean model. We have implemented a Lagrangian method which takes into account the four dimensional (space and time) nature of the observations: RMS errors with assimilation are twice lower than without and the main patterns of the fluid flow are well identified at each vertical level, although the floats drift at a single determined level.

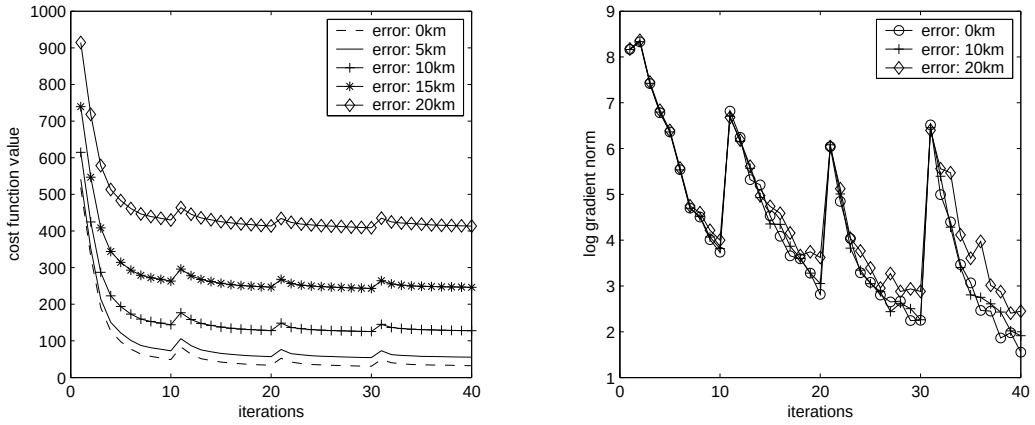

Figure 12: Evolution of the cost function and its gradient's norm during the assimilation of noisy data. On the left, evolution of the cost function for experiments with noise amplitude from 1km to 20km. For reference, the cost for experiment without noise (0km) is also plotted. On the right, evolution of the gradient norm, shown on a logarithmic scale, for experiments with and without noise in data.

We have tested the sensitivity of our method to the characteristics of the dataset. It is very sensitive to the vertical drift level, and the best results are obtained for intermediates ones, especially level 4 (around 1000 meters depth). It is also very sensitive to the number of floats, but the more is not the better, it seems useless to launch more than 1 000 floats in our configuration. It is very robust with respect to the increase of the time-sampling period, up to ten days.
We have compared our Lagrangian method to the Eulerian one, which consists in interpreting Lagrangian data as velocity information. When the time-sampling period of the observations is one day or less, the Eulerian method performs slightly better, but the transfer of information to lower levels is better achieved by the Lagrangian one. When this period is larger than two or three days, the Lagrangian method performs much better than the Eulerian one.
We also studied the impact of errors on observation: the reconstruction of the velocities is well achieved even with a large noise in data.
Also the performances of this method have been assessed in the framework

of the twin experiments approach. The next step would be to use real data and to deal with problems such as trajectories modelling and model error.

Acknowledgments

The author thanks Anthony Weaver and LODYC for the source code of OPAVAR 8.1.
Numerical computations were performed on the NEC SX5 vector computer at IDRIS.
This work is supported by French project Mercator.

References

References

[1] Assenbaum M and Reverdin G 2005 Near real-time analysis of the mesoscale circulation during the POMME experiment *Deep-sea Research Part I* **52** 1345–1373

[2] ——2005 Private communication

[3] Courtier P, Thépaut J-N and Hollingsworth A 1994 A strategy for operational implementation of 4D-Var, using an incremental approach *Quart. J. Roy. Meteor. Soc* **120** 1367–87

[4] De Mey P 1997 Data assimilation at the oceanic mesoscale: a review *Journal of the Meteorological society of Japan* **75** 1B 415–27

[5] Forget G 2004 4D-Var assimilation of Argo profiles applied to North Atlantic Ocean climate monitoring University of Bretagne Occidentale Brest France

[6] Ghil M 1989 Meteorological data assimilation for oceanographers Part I: description and theoretical framework *Dyn. Atmos. Oceans* **13** 171–218

[7] ——and Manalotte-Rizzoli P 1991 Data assimilation in meteorology and oceanography *Adv. Geophys.* **23** 141–265

[8] ——, Ide K, Bennett A, Courtier P, Kimoto M, Nagata M, Saiki M and Sato N 1997 Data assimilation in meteorology and oceanography: Theory and Practice *Meteorological Society of Japan*

[9] Giering R and Kaminski T 1998 Recipes for Adjoint Code Construction *ACM Trans. On Math. Software* **24** 4 437–74

[10] Holland WR 1978 The Role of Mesoscale Eddies in the General Circulation of the Ocean – Numerical Experiments Using a Wind-Driven Quasi-Geostrophic Model *Journal of Physical Oceanography* **8** 363–392

[11] Ide K, Courtier P, Ghil M and Lorenc AC 1997 Unified notation for data assimilation: operational, sequential and variational *J. Meteor. Soc. Japan* **75** 1B 181–9

[12] ——, Kuznetsov L and Jones C 2002 Lagrangian data assimilation for point-vortex system *J. Turbulence* **3** 053

[13] Jazwinski AH 1970 Stochastic processes and filtering theory *Applied Mathematical Sciences* **64** Academic Press

[14] Kamachi M and O'Brien J 1995 Continuous data assimilation of drifting buoy trajectory into an equatorial Pacific Ocean model *Journal of Marine Systems* **6** 159–78

[15] Kuznetsov L, Ide K and Jones C 2003 A method for assimilation of Lagrangian data *Mon. Wea. Rev.* **131**(10) 2247–2260

[16] Le Dimet F X, Navon I M and Daescu D N 2002 Second-Order Information in Data Assimilation *Journal: Monthly Weather Review* **130** 3 629–648

[17] ——and Talagrand O 1986 Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects *Tellus* A **38** 97

[18] Lions J L 1968 *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles* (Paris: Dunod Gauthier-Villars)

[19] ——, Temam R and Wang S 1992 On the equations of the large-scale ocean *Nonlinearity* **5** 5 1007–53

[20] Madec G, Delecluse P, Imbard M and Levy C 1998 OPA8.1 ocean general circulation model reference manual *Notes du pole de Modelisation de l'IPSL* **11**

[21] Mead J 2005 Assimilation of simulated float data in Lagrangian coordinates *Ocean Modeling* **8** 369–94

[22] Molcard A, Piterbarg L, Griffa A, Özgökmen T and Mariano A 2003 Assimilation of drifter observations for the reconstruction of the Eulerian circulation field *J. Geophys. Res. Oceans* **108** C03 3056 1–21

[23] Ngodock H E 1996 Data assimilation and sensitivity analysis: application to ocean circulation, PhD thesis, University Joseph Fourier Grenoble France

[24] Ozgökmen T, Molcard A, Chin T, Piterbarg L and Griffa A 2003 Assimilation of drifter observation in primitive equation models of mid-latitude ocean circulation *J. Geophys. Res. Oceans* **108** C07 3238 31 1–31 17

[25] Penenko V V and Obraztsov N N 1976 A variational initialization method for the fields of the meteorological elements *Sov. Meteorol. Hydrol.* **11** 1–11

[26] Reynaud T 2005 Private communication

[27] Salman H, Kuznetsov L, Jones C and Ide K 2005 A method for assimilating Lagrangian data into a shallow-water equation ocean model *Mon. Wea. Rev.* to appear

[28] Sheinbaum J and Anderson D L T 1990 Variational assimilation of XBT data Part I *J. Phys. Oceanogr.* **20** 672–88

[29] Talagrand O 1991 The use of adjoint equations in numerical modeling of the atmospheric circulation *Automatic differentiation of algorithms, Proc. 1st SIAM Workshop, Beckenridge/ CO (USA)* 169–180

[30] ——and Courtier P 1987 Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory *Quarterly Journal of the Royal Meteorological Society* **113** 478 1311–28

[31] Temam R and Ziane M 2004 Some mathematical problems in geophysical fluid dynamics *Handbook of Mathematical Fluid Dynamics 3* (Friedlander S and Serre D Editors Elsevier)

[32] Thacker W C and Long R B 1988 Fitting dynamics to data *J. Geophys. Res.* **93** 1227–40

[33] Wang Z, Navon I M, Le Dimet F X and Zou X 1992 The second order adjoint analysis: Theory and applications *Meteorology and Atmospheric Physics (Historical Archive)* **50** (1-3) 3–20

[34] Weaver A T and Courtier P 2001 Correlation modeling on the sphere using a generalized diffusion equation *Q. J. R. Meteorol. Soc.* **127** 1815–46

[35] ——, Vialard J and Anderson D L T 2003 Three- and Four-dimensional variational assimilation with an ocean general circulation model of the tropical Pacific Ocean Part I: formulation, internal diagnostics and consistency checks *Mon. Wea. Rev.* **131** 1360–78